

# Estimating Species Trees from Quartet Gene Tree Distributions under the Coalescent Model

Martin Kreidl

August 9, 2011

## Abstract

In this article we propose a new method, which we name ‘quartet neighbor joining’, or ‘quartet-NJ’, to infer an unrooted species tree on a given set of taxa  $T$  from empirical distributions of unrooted quartet gene trees on all four-taxon subsets of  $T$ . In particular, quartet-NJ can be used to estimate a species tree on  $T$  from distributions of gene trees on  $T$ . The quartet-NJ algorithm is conceptually very similar to classical neighbor joining, and its statistical consistency under the multispecies coalescent model is proven by a variant of the classical ‘cherry picking’-theorem. In order to demonstrate the suitability of quartet-NJ, coalescent processes on two different species trees (on five resp. nine taxa) were simulated, and quartet-NJ was applied to the simulated gene tree distributions. Further, quartet-NJ was applied to quartet distributions obtained from multiple sequence alignments of 28 proteins of nine prokaryotes.

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>The multispecies coalescent model</b>	<b>4</b>
2.1	Overview . . . . .	4
2.2	Quartet distributions . . . . .	6
<b>3</b>	<b>A cherry picking theorem</b>	<b>7</b>
3.1	Depth of a pair and cherry picking . . . . .	7
3.2	Statistical consistency of cherry picking . . . . .	11
3.3	Minimal probability for incomplete lineage sorting . . . . .	12

<b>4</b>	<b>Quartet neighbor joining algorithms</b>	<b>13</b>
4.1	A naive neighbor joining algorithm . . . . .	14
4.2	Perturbing weights and quartet-NJ . . . . .	15
4.3	Reduction of computational complexity . . . . .	18
<b>5</b>	<b>Application and simulations</b>	<b>19</b>
5.1	Application to a prokaryote data set . . . . .	19
5.2	Simulations with Mesquite . . . . .	22
	<b>References</b>	<b>25</b>

# 1 Introduction

As the amount of available sequence data from genomes or proteoms rapidly grows, one is confronted with the problem that, for a given set of taxa, tree topologies relating homologous sequences of these taxa can (and in practice often will) differ from the each other, depending on which locus in the genome or which protein is used for constructing the tree. Possible reasons for discordance among gene trees and the species phylogeny include horizontal gene transfer, gene duplication/loss, or incomplete lineage sorting (deep coalescences). Thus one is confronted with the task to determine the phylogeny of the taxa (the ‘species tree’) from a set of possibly discordant gene trees (Maddison [8], and Maddison and Knowles [9]).

Several methods have been proposed and are used to infer a species tree from a set of possibly discordant gene trees:

(1) Declaring the most frequently observed gene tree to be the true species tree was shown to be statistically inconsistent under the multispecies coalescent model by Degnan and Rosenberg [4], in cases where branch lengths of the underlying species tree are sufficiently small.

(2) A popular approach for species tree reconstruction is by concatenating multiple alignments from several loci to one large multiple alignment and construct a single ‘gene tree’ from this multiple alignment by standard methods. This approach was pursued e.g. by Teichmann and Mitchison [17] and Cicarelli et al. [2]. However, also this method was shown to be inconsistent under the multispecies coalescent by Degnan and Kubatko [5], if branch lengths on the species tree are sufficiently short.

(3) Similarly, the concept of minimizing the number of ‘deep coalescences’ (Maddison [8]) was recently shown to be statistically inconsistent under the multispecies coalescent model by Than and Rosenberg [19].

(4) On the other hand, it is well known from coalescent theory that the probability distribution of gene trees, or even the distributions rooted gene triplet trees, on a species tree uniquely determine the species tree if incomplete lineage sorting is assumed to be the only source of gene tree discordance. Similarly (see Allman et al. [1]), *the distributions of unrooted quartet gene trees identify the unrooted species tree topology*. However, as soon as triplet/quartet gene trees are inferred from experimental data, their distributions will differ from the theoretical ones, and this may lead to conflicts among the inferred triplets/quartets on the hypothetical species tree, a problem which is not straight forward to resolve (‘quartet-puzzeling’ could be an approach to resolve this, see Strimmer and von Haeseler [16]). Also direct maximum likelihood calculations using gene tree distributions are very problematic due to the enormous number of possible gene trees on a given set of taxa.

(5) Recently Liu and Yu [7] have proposed to use the ‘average internode distance’ on gene trees as a measure of distance between taxa on species trees and apply classical neighbor joining to these distance data. They prove that this reconstructs the underlying species tree in a statistically consistent way.

In the present paper we propose a method to overcome many of the difficulties discussed above, and in particular to make the above mentioned result by Allman et al. [1] accessible in practice. That is, we describe a polynomial time algorithm (in the number of taxa), which uses empirical distributions of unrooted quartet gene trees as an input, to estimate the unrooted topology of the underlying species tree in a statistically consistent way. Due to the conceptual similarity to classical neighbor joining we call this algorithm ‘quartet neighbor joining’, or briefly ‘quartet-NJ’.

The fact that quartet-NJ uses quartet distributions as an input makes it flexible in practice: Quartet gene tree distributions can be obtained directly from sequence data (as is described in the application to a prokaryote data set in Section 5 of this paper), or e.g. from gene tree distributions, which is the type of input required by Liu and Yu’s ‘neighbor joining for species trees’ [7]. Also, quartet-NJ is naturally capable of dealing with multiple

lineages per locus and taxon.

The paper is organized as follows: After a brief review of the multispecies coalescent model and distributions of quartet gene trees in Section 2, we investigate in Section 3 how to identify cherries on a species tree. To this end we show how to assign ‘weights’ to unrooted quartet trees on the set of taxa, using the distributions of quartet gene trees, and how to define a ‘depth’ for each pair of taxa. In analogy to classical cherry picking we prove that under the multispecies coalescent model any pair of taxa with maximal depth is a cherry on the species tree. Moreover, we give an interpretation of this theorem by the concept of ‘minimal probability for incomplete lineage sorting’, in analogy to the concept of minimal evolution underlying classical neighbor joining.

In Section 4 we translate our cherry picking theorem into a neighbor joining-like procedure (‘quartet-NJ’) and prove that it reproduces the true unrooted species tree topology as the input quartet distributions tend to the theoretical quartet distributions. In other words, quartet-NJ is statistically consistent.

Finally, in Section 5 we apply the quartet-NJ algorithm to data from coalescence simulations, as well as to a set of multiple alignments from nine prokaryotes, in order to demonstrate the suitability of quartet-NJ. In both situations we consider only one lineage per locus and taxon.

## 2 The multispecies coalescent model

We are going to present briefly the multispecies coalescent, which models gene (or protein) tree evolution within a fixed species tree. The material in this section is neither new nor very deep. Rather this section is to be considered as a reminder for the reader on the relevant definitions, as well as a suitable place to fix language and notation for the rest of the paper. For a more detailed introduction to the multispecies coalescent model we refer e.g. to Allman et al. [1].

### 2.1 Overview

Let us consider a set  $T$  of  $n$  taxa (e.g. species) and a rooted phylogenetic (that is: metric with strictly positive edge lengths and each internal node is trivalent) tree  $S$  on the taxon set  $T$ . The tree  $S$  is assumed to depict the

‘true’ evolutionary relationship among the taxa in  $T$ , and the lengths of its internal edges are measured in coalescence units (see below). This tree is commonly called the *species tree* on the taxon set  $T$ .

It is a fundamental problem in phylogenetics to determine the topology of this tree. Molecular methods, however, usually produce evolutionary trees for single loci within the genome of the relevant species, and these *gene trees* will usually differ topologically from the species tree (Degnan and Rosenberg [4]). This has several biological reasons, like horizontal gene transfer, gene duplication/loss or incomplete lineage sorting. The latter describes the phenomenon that gene lineages diverge long before the population actually splits into two separate species, and in particular gene lineages may separate in a different order than the species do.

The multispecies coalescent model describes the probability for each rooted tree topology to occur as the topology of a gene tree, under the assumption that incomplete lineage sorting is the only reason for gene tree discordance.

Notation: We adopt the common practice to denote taxa (species) by lower case letters, while we denote genes (or more general: loci) in their genome by capital letters. E.g. if we consider three taxa  $a, b, c \in T$ , the letters  $A, B, C$  will denote a particular locus sampled from the respective taxa. (By abuse of notation, we will identify the leaf sets of both species *and* gene trees with  $T$ ).

Considering only a single internal branch of the species tree, and two gene lineages within this branch going backwards in time, we find that the probability that the two lineages coalesce to a single lineage within time  $\tau$  is given by

$$P = 1 - \exp(-\tau)$$

where  $\tau$  is in *coalescence units* (that is number of generations represented by the internal branch divided by the total number of allele of the locus of interest, present in the population). In particular, if the branch on the species tree has length  $d$ , the probability that two lineages coalesce within this branch is  $1 - \exp(-d)$ .

From this, Nei [12] derives the following probabilities under the multi-species coalescent model for a three taxon tree. Denote the taxa by  $a, b, c$ , the corresponding loci by  $A, B, C$ , and assume that the species tree has the

topology  $((a, b), c)$  with internal edge length  $d > 0$ . Then the probability that a gene tree sampled from these taxa has the topology  $((A, B), C)$  is equal to

$$1 - \frac{2}{3} \exp(-d),$$

while the other two topologies are observed with probability  $\frac{1}{3} \exp(-d)$ .

In [3] Degnan and Salter show how to calculate the probabilities of a gene tree in a fixed species tree on an arbitrary taxon set. It turns out that these probabilities are polynomials in the unknowns  $\exp(-d_E)$ , where  $d_E$  denotes the length of the edge  $E$  on the species tree. In particular, the multispecies coalescent model is an algebraic statistical model.

## 2.2 Quartet distributions

In the following we consider four taxon species trees on the taxon set  $T = \{a, b, c, d\}$ , and we use Newick notation to specify (rooted) species trees. For instance  $S_1 = (((a, b) : x, c) : y, d)$  denotes the caterpillar tree with edge lengths  $x$  and  $y$  in coalescent units. Moreover, unrooted quartets are denoted in the format  $(AB, CD)$ , meaning the quartet gene tree which has cherries  $(AB)$  and  $(CD)$ .

Up to permutation of leaf labels there is only one additional tree topology on four taxa, namely the balanced tree  $S_2 = ((a, b) : x, (c, d) : y)$ . For both of these two species tree topologies it is not hard to calculate the probabilities of the tree possible unrooted quartet gene trees on  $T$ , and one obtains

**Lemma 1** (Allman et al. [1]). *For both species tree topologies  $S = S_1$  and  $S = S_2$  the probabilities of unrooted quartet gene trees on  $T$  have the same form and are given by the formulas*

$$P(AB, CD) = 1 - \frac{2}{3} \exp(-x - y) > \frac{1}{3}, \quad (2.1)$$

$$P(AC, BD) = \frac{1}{3} \exp(-x - y) < \frac{1}{3}, \quad (2.2)$$

$$P(AD, BC) = \frac{1}{3} \exp(-x - y) < \frac{1}{3}. \quad (2.3)$$

In particular, the gene quartet distribution determines the *unrooted* species tree topology, but not the position of the root (Allman et al.). The

sum of the internal edge lengths of the rooted species tree (resp. the length of the internal edge of the unrooted species tree) is given by the formula

$$d(ab, cd) := d = x + y = -\log\left(\frac{3}{2}(1 - P(AB, CD))\right). \quad (2.4)$$

Returning to the study of the multispecies coalescent on species trees on arbitrary taxon sets, we deduce from the above that the quartets displayed on the species tree  $S$  on the taxon set  $T$  are exactly those which appear with a probability bigger than  $\frac{1}{3}$  for any sampled locus. Hence, the distributions of quartets determine uniquely the quartet subtrees which are displayed by the true species tree  $S$ , and hence determine the unrooted topology of  $S$  (see Allman et al. [1]).

In the following two sections we describe a neighbor joining algorithm which makes this theoretical insight applicable in practice, meaning that it yields a method to estimate unrooted species trees from (empirical) distributions of quartet gene trees, which is statistically consistent under the multispecies coalescent model. Statistical consistency of this algorithm will follow from the ‘cherry picking theorem’ below.

### 3 A cherry picking theorem

Here we give a precise criterion, using theoretical distributions of quartet gene trees under the multispecies coalescent, to determine which pairs of taxa on the species tree are cherries. This criterion can as well be applied to estimated quartet distributions and thus enables us to recursively construct a species tree estimate from observed gene quartet frequencies in Section 4.

#### 3.1 Depth of a pair and cherry picking

As always, we consider a fixed species tree  $S$  on the taxon set  $T$ , and we denote by  $P(IJ, KL) := P_S(IJ, KL)$  the probability that a gene tree on  $T$  displays the gene quartet tree  $(IJ, KL)$ .

Recall that by lower case letters  $i, j, l, k$  we denote the species from which the genes  $I, J, K, L$  are sampled. Regardless of whether  $S$  contains the (species-) quartet  $(ij, kl)$ , we can attach to it the following numbers:

**Definition 2** (Weight of a quartet). *The weight of the quartet  $(ij, kl)$  is*

defined as

$$w(ij, kl) = -\log\left(\frac{3}{2}(P(IK, JL) + P(IL, JK))\right). \quad (3.1)$$

If the taxa  $i, j, k, l$  are not pairwise distinct, then we set  $w(ij, kl) = 0$ .

Of course, if  $i, j, k, l$  are pairwise distinct we may equivalently write

$$w(ij, kl) = -\log\left(\frac{3}{2}(1 - P(IJ, KL))\right).$$

**Lemma 3.** *If the quartet  $(ij, kl)$  is displayed on the species tree  $S$ , then the weight  $w(ij, kl)$  is precisely the length of the interior branch of the quartet  $(ij, kl)$ . Moreover, the other two weights,  $w(ik, jl)$  and  $w(il, jk)$ , are less than zero.*

*Proof.* The proof is immediate using equation (2.4): If the quartet tree  $(ij, kl)$  is displayed on the species tree  $S$ , then

$$\begin{aligned} w(ij, kl) &= -\log\left(\frac{3}{2}(P(IK, JL) + P(IL, JK))\right) = \\ &= -\log\left(\frac{3}{2}\frac{2}{3}\exp(-d(ij, kl))\right) = d(ij, kl). \end{aligned} \quad (3.2)$$

Otherwise, if  $(ij, kl)$  is not displayed on  $S$ , then one of the other two quartet trees on  $\{i, j, k, l\}$  is displayed, say  $(ik, jl)$ , and we have

$$\begin{aligned} w(ij, kl) &= -\log\left(\frac{3}{2}(P(IK, JL) + P(IL, JK))\right) = \\ &= -\log\left(\frac{3}{2}\left(1 - \frac{1}{3}\exp(-d(ik, jl))\right)\right) < -\log\left(\frac{3}{2}\frac{2}{3}\right) = 0, \end{aligned} \quad (3.3)$$

since  $d(ik, jl) > 0$ . □

This little lemma motivates the following definition and a cherry picking theorem which is formulated and proved in close analogy to the ‘classical’ one (see Saito and Nei [13] for the original publication, or Pachter and Sturmfels [14] for a presentation analogous to ours).

**Definition 4** (Depth of a pair). *For any pair of taxa  $i, j \in T$  we define the depth of  $(i, j)$  to be the number*

$$f(i, j) = \sum_{k, l \in T} w(ij, kl). \quad (3.4)$$



(Recall that if  $\{i, j\} \cap \{k, l\} \neq \emptyset$  then  $w(ij, kl) = 0$ .)

**Theorem 5** (Cherry Picking). *If a pair of taxa  $i, j \in T$  has maximal depth  $f(i, j)$ , then  $(ij)$  is a cherry on the species tree  $S$ .*

As mentioned above, the proof of this theorem is parallel to the proof of the classical cherry picking result as presented in [14]. We state it here for sake of completeness of our exposition.

*Proof.* As a first step, we introduce the auxiliary values

$$v(ij, kl) := \begin{cases} w(ij, kl) & \text{if } (ij, kl) \text{ is displayed on } S, \\ 0 & \text{else,} \end{cases} \quad (3.5)$$

for every four taxon subset  $\{i, j, k, l\} \subset T$ , as well as

$$g(i, j) := \sum_{k, l} v(ij, kl) \quad (3.6)$$

for every pair of taxa. Obviously, if  $(ij)$  is a cherry on  $S$ , then  $w(ij, kl) = v(ij, kl)$  for all  $k, l \in T$ , and hence also  $g(i, j) = f(i, j)$ . In general, for any pair of taxa  $(i, j)$  one has  $g(i, j) \geq f(i, j)$  by Lemma 3.

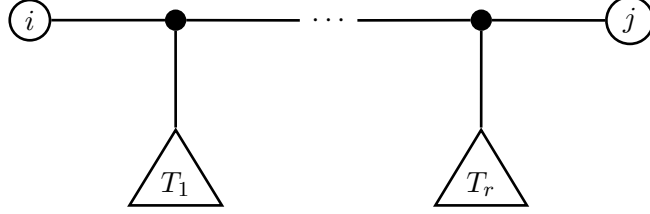
We will now assume that the pair  $(i, j)$  does *not* form a cherry on the species tree  $S$  and prove that in this case there exists a cherry  $(pq)$  on  $S$  such that the following inequalities hold:

$$f(p, q) = g(p, q) > g(i, j) \geq f(i, j). \quad (3.7)$$

From this we deduce the claim of the theorem: If  $(ij)$  is not a cherry on  $S$ , then it does not have maximal depth.

*Proof of the claim.* We have to find a cherry  $(pq)$  on  $S$  such that indeed  $g(p, q) > g(i, j)$  holds. As we assume  $i$  and  $j$  not to form a cherry on  $S$ , the unique path connecting  $i$  and  $j$  crosses at least two interior nodes (symbolized as black dots in figure 1. We denote the number of internal nodes on this path by  $r \geq 2$ . To each  $s = 1, \dots, r$  a rooted binary tree  $T_j$  is attached (see Figure 1). By interchanging  $i$  and  $j$ , if necessary, we may assume that the number of taxa in  $T_1$  is less or equal than the number of taxa in  $T_r$ .

Figure 1: The nodes  $i$  and  $j$  do not form a cherry.



First we consider the case that the subtree  $T_1$  consists of a single leaf  $i'$ . In this case the pair  $(i, i')$  is a cherry on  $S$ , and it is easy to see that indeed  $g(i, i') > g(i, j)$ .

Second, if  $T_1$  has more than one leaf, let choose a cherry  $(pq)$  in  $T_1$  (every rooted binary tree with more than one leaf has at least one cherry!). Now we decompose the difference  $g(p, q) - g(i, j)$  into six sums which will be treated separately:

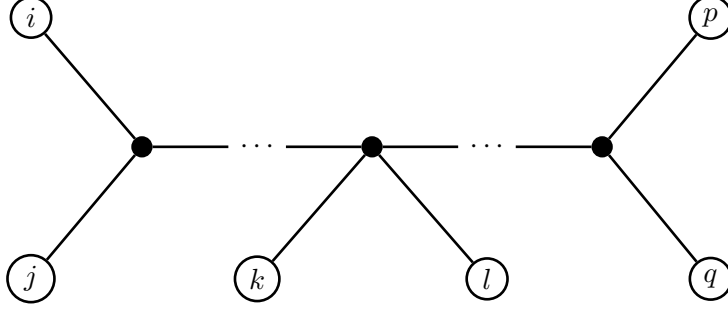
$$g(p, q) - g(i, j) = S_1 + \dots + S_5, \quad (3.8)$$

where

$$\begin{aligned} S_1 &= \sum_{k, l \in T_m, m=2, \dots, r} (v(pq, kl) - v(ij, kl)), \\ S_2 &= \sum_{k \in T_m \setminus \{p, q\}, l \in T_{m'} \setminus \{p, q\}, m, m'=1, \dots, r, m \neq m'} (v(pq, kl) - v(ij, kl)), \\ S_3 &= \sum_{k, l \in T_1 \setminus \{p, q\}} (v(pq, kl) - v(ij, kl)), \\ S_4 &= \sum_{k \in T_m, m=2, \dots, r} (v(pq, ki) + v(pq, kj) - v(ij, kq) - v(ij, kp)), \\ S_5 &= \sum_{k \in T_1 \setminus \{p, q\}} (v(pq, ki) + v(pq, kj) - v(ij, kq) - v(ij, kp)) + \\ &\quad + (-v(pq, ij) + v(ij, pq)). \end{aligned}$$

Inspection of the tree in Figure 1 immediately shows that each summand in  $S_1$  is positive, and more precisely, greater or equal to  $d(ij, pq)$ . Hence  $S_1 > \binom{|T_r|}{2} d(ij, pq)$ . In  $S_2$ , the expressions  $v(ij, kl)$  all vanish, hence this sum is positive. The third sum  $S_3$  might be negative - but not too negative: The situation is illustrated by Figure 2 (which arises from the tree in Figure 1 by deleting the irrelevant subtrees  $T_2, \dots, T_r$ ), whose inspection shows that

Figure 2: Part of the tree: leaves  $i, j$  and subtree  $T_1$  with cherry  $(pq)$  and leaves  $k, l$ .



each for each summand  $v(pq, kl) - v(ij, kl)$  we have

$$v(pq, kl) - v(ij, kl) = d(pq, kl) - d(ij, kl) \geq -d(ij, pq).$$

Thus we see  $S_3 \geq -\binom{|T_1|-2}{2}d(ij, pq) > -\binom{|T_r|}{2}d(ij, pq)$ , whence  $S_1 + S_2 + S_3$  is positive.

Now we treat the fourth sum: If  $k$  is a node in one of the subtrees  $T_2, \dots, T_r$ , then  $v(ij, kq) = v(ij, kp) = 0$ , so  $S_4$  is non-negative. More precisely we have  $S_4 \geq 2\binom{|T_r|}{2}v(pq, ij)$ . On the other hand, the sum  $S_5$  is greater or equal  $-2\binom{|T_1|-2}{2}v(ij, pq) \geq -2\binom{|T_r|}{2}v(pq, ij)$ , whence also  $S_4 + S_5$  is positive. In total we have found that in any case  $g(p, q) > g(i, j)$ , which proves our claim.  $\square$

As explained above, this suffices to prove the cherry picking theorem.  $\square$

### 3.2 Statistical consistency of cherry picking

In practical applications one will not know the precise probability of each quartet gene tree  $(AB, CD)$ . Hence one has to use e.g. relative frequencies as estimates. Let us denote by  $r(AB, CD)$  the (experimentally determined) relative frequency of the gene quartet  $(AB, CD)$ . In analogy to Definitions 2 and 4 we make the following

**Definition 6.** *The empirical weight of the quartet  $(ij, kl)$  of taxa is defined as*

$$\tilde{w}(ij, kl) = -\log\left(\frac{3}{2}(r(IK, JL) + r(IL, JK))\right), \quad (3.9)$$

Moreover, the empirical depth of a pair  $(i, j)$  of taxa is defined as

$$\tilde{f}(i, j) = \sum_{k, l \in T} \tilde{w}(ij, kl). \quad (3.10)$$

Under the multispecies coalescent model, the relative frequency  $r(IJ, KL)$  of a quartet gene tree  $(IJ, KL)$  is a statistically consistent estimator for its probability  $P(IJ, KL)$ . Hence the empirical depth  $\tilde{f}(i, j)$  of a pair of taxa  $i, j$  is a statistically consistent estimator for  $f(i, j)$ . In particular this means: inferring a pair  $(i, j)$  of taxa, where  $\tilde{f}$  is maximal, as a cherry on the species tree is statistically consistent. More precisely, we have

**Corollary 7.** *Let  $N$  be the number of loci sampled from each taxon in  $T$  and denote by  $C_N$  the set of pairs of taxa at which  $\tilde{f}$  attains its maximum. If the number of genes  $N$  tends to infinity, then the probability that  $C_N$  contains a pair which is not a cherry on the true species tree approaches zero.*

*Proof.* As there are only finitely many taxa in  $T$ , there is a strictly positive difference between the maximum value of  $f$  on  $T^2$  and its second biggest value. Since moreover  $f$  and  $\tilde{f}$  are continuous, there exists an  $\delta > 0$  such that whenever  $|r(IJ, KL) - P(IJ, KL)| < \delta$ ,  $\tilde{f}(p, q)$  is maximal if and only if  $f(p, q)$  is maximal, and the claim follows from Theorem 5. But the probability that  $|r(IJ, KL) - P(IJ, KL)| < \delta$  approaches 1 as  $N$  grows, for any choice of  $\delta$ .  $\square$

### 3.3 Minimal probability for incomplete lineage sorting

Recall that classical neighbor joining is a greedy algorithm which in each ‘cherry picking’ step declares a pair of taxa to be neighbors if this minimizes the sum of branch lengths in the refined tree resulting from this step (Saito and Nei [13]). In other words, classical cherry picking and neighbor joining is guided by the principle of *minimal evolution*. It turns out that our cherry picking result in Theorem 5 can be interpreted in a similar fashion.

Let us make the following ad-hoc definition.

**Definition 8.** *Let  $X_1, \dots, X_n$  be independent random variables with values in  $\{0, 1\}$ , and let  $p_i$  be the probability of  $\{X_i = 1\}$  for each  $i = 1, \dots, n$ . We call the geometric mean*

$$\bar{p} = \sqrt[n]{p_1 \cdots p_n} \quad (3.11)$$

the average probability of the random experiments  $X_1, \dots, X_n$  giving the result 1.

Let us now consider what the cherry picking Theorem 5 does with a set of taxa  $T$ , initially arranged as a star-like tree (see Figure 3). Consider two taxa  $i \neq j \in T$  fixed and let  $q(ij, kl) = P(IK, JL) + P(IL, JK)$ . On a species tree on  $T$  which displays the cherry  $(ij)$ , this is the probability that a gene quartet tree sampled from the taxa  $i, j, k, l$  differs from the species quartet tree  $(ij, kl)$ . Let

$$\bar{q}(i, j) = \binom{n-2}{2} \sqrt{\prod_{k, l} q(ij, kl)}$$

be the average probability of discordance of a sampled gene quartet tree with the corresponding quartet on the species tree, for a set of four taxa containing  $i$  and  $j$ . We will also call this number the ‘average probability of incomplete lineage sorting’ for quartets containing the taxa  $i$  and  $j$ . With this terminology, the cherry picking Theorem 5 can be phrased as follows.

**Theorem 9.** *A pair of taxa  $i, j \in T$  is a cherry on the true species tree  $S$  if it minimizes the average probability for incomplete lineage sorting for quartets containing the taxa  $i$  and  $j$ .*

*Proof.* Using Theorem 5 we only have to show that  $\bar{q}(i, j)$  is minimal if and only if  $f(i, j)$  is maximal. But this follows from the fact that

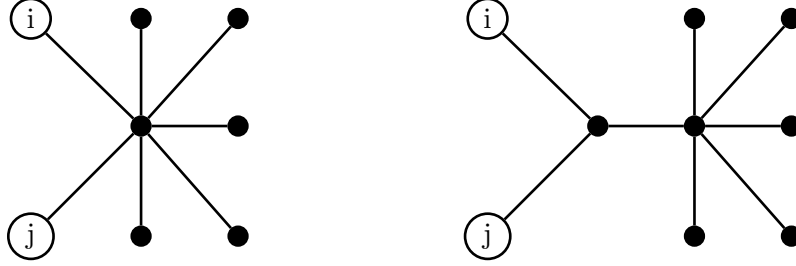
$$f(i, j) = -\log \bar{q}(i, j) + C,$$

where  $C$  is a constant independent of  $i, j$ . □

## 4 Quartet neighbor joining algorithms

In this section we discuss how the above cherry picking result can be used to design an algorithm which estimates a species tree from (observed) gene quartet tree frequencies.

Figure 3: One cherry picking step.



#### 4.1 A naive neighbor joining algorithm

The first version of our neighbor joining algorithm reconstructs the underlying species tree from the (theoretical) gene quartet tree distributions.

**Algorithm 10.** *Input:* A set of taxa  $T = \{t_1, \dots, t_n\}$  containing at least four elements, and for each quartet  $Q = (ab, cd)$  with  $a, b, c, d \in T$  the probability  $r(Q) := P_S(Q)$  of the quartet  $Q$  under the multispecies coalescent model on the species tree  $S$ .

*Output:* The species tree  $S$ .

**Step 0.** Set  $S$  to be the graph with vertex set  $T$  and with empty edge set.

**Step 1.** If  $|T| \leq 3$  go to step 4.

**Step 2.** For each unordered pair  $\{i, j\} \subset T$  calculate the depth  $f(i, j)$ . Let  $i, j \in T$  be a pair of taxa with maximal depth. Add a node  $x$  to  $S$  and draw edges from  $i$  to  $x$  and from  $j$  to  $x$ . Replace  $T$  by  $T \cup \{x\} \setminus \{i, j\}$ .

**Step 3.** For each quartet  $(xa, bc)$  with  $a, b, c \in T \setminus \{x\}$  set

$$r'(xa, bc) = \frac{1}{2}(r(ia, bc) + r(ja, bc)),$$

and for all quartets  $(ab, cd)$  which do not contain  $x$  set  $r'(ab, cd) = r(ab, cd)$ . Replace  $r$  by  $r'$ . Go to step 1.

**Step 4.** If  $|T| = 3$ , then add a vertex to  $S$  and draw edges from this new vertex to the three elements of  $T \subset \text{vertices}(S)$ .  $\square$

**Remark 11.** Each calculation of  $f(i, j)$  requires  $O(|T|^2)$  logarithms and additions. Repeating this for every pair  $i, j$  of taxa gives a computational complexity of  $O(|T|^4)$  for step 2. This is also the complexity of one iteration

of the algorithm. Since each step will identify exactly one cherry, we need  $O(|T|^5)$  iterations to reconstruct  $S$ , which gives a total complexity of  $O(|T|^5)$  for Algorithm 10. It is rather straightforward to improve step 2 so that the algorithm requires only  $O(|T|^4)$  arithmetic operations (see Subsection 4.3).

Using the result of Theorem 9, we may summarize: *With Algorithm 10 we have described a greedy algorithm which infers in each step the cherry which requires minimal incomplete lineage sorting.* Further, the fact that cherry picking is statistically consistent implies that also species tree estimation by Algorithm 10 is statistically consistent:

**Corollary 12** (Statistical consistency of Algorithm 10). *Let  $N$  be the number of loci sampled from each taxon in  $T$  and denote by  $S_N$  the estimate for the species tree produced by Algorithm 10. If the number of genes  $N$  goes to infinity, then the probability that  $S_N$  equals the true species tree  $S$  approaches 1.*

*Proof.* This follows by a repeated application of Corollary 7. □

In practical applications two problems may occur using Algorithm 10: First, there might be several non-disjoint pairs of taxa with maximal depth. This could be resolved by choosing one of these pairs randomly. A more serious problem occurs when many of the empirical gene quartet distributions are 0. This will be discussed in the following subsection.

## 4.2 Perturbing weights and quartet-NJ

Algorithm 10 works well if one uses as input the theoretical probabilities for quartet gene trees. It also works well, if one uses relative quartet frequencies which are ‘very close’ to the probabilities  $P(IJ, KL)$  (and in particular non-zero). In other cases, the result might be problematic, as we are going to explain now.

Imagine that for some reason (e.g. very long branch lengths on the species tree) the observed quartet gene trees reflect very precisely the topology of the species tree. This might mean in the extreme case that for every four-taxon subset  $\{I, J, K, L\}$  one observes only this very quartet gene tree which is displayed also by the (unknown) species tree. In some sense this situation should be optimal, since the observed gene quartet trees fit together without conflict and thus yield an unambiguous estimate for the species tree.

However, let us run Algorithm 10 with the empirical depth function  $\tilde{f}$  in place of  $f$  and consider any pair  $i, j$  of taxa. Regardless of the choice of  $i$  and  $j$ , there exist  $k, l \in T$  such that  $(ij, kl)$  is a quartet on the species tree, and hence by our assumption  $r(IJ, KL) = 1$ . Calculating the empirical depth of  $(i, j)$  yields

$$\tilde{f}(i, j) = \sum_{k, l \in T} -\log\left(\frac{3}{2}(1 - r(IJ, KL))\right) = +\infty. \quad (4.1)$$

Thus *every* pair  $(i, j)$  maximizes  $\tilde{f}$ , and so in each iteration of Algorithm 10 the choice of a cherry is completely arbitrary. So this procedure fails in a situation, which has to be considered the ‘easiest’ possible in some obvious sense.

A possible solution to this problem is to perturb the arguments in the logarithms in equation (4.1) by a small number  $\epsilon > 0$ . In other words, we fix  $0 \leq \epsilon \ll 1$  and calculate, for each pair of taxa  $i, j$ , the ‘perturbed depths’

$$f_\epsilon(i, j) = \sum_{k, l \in T} -\log\left(\frac{3}{2}((1 + \epsilon) - P(IJ, KL))\right) < \infty, \quad (4.2)$$

$$\tilde{f}_\epsilon(i, j) = \sum_{k, l \in T} -\log\left(\frac{3}{2}((1 + \epsilon) - r(IJ, KL))\right) < \infty. \quad (4.3)$$

Obviously, for  $\epsilon \rightarrow 0$  we have  $f_\epsilon(i, j) \rightarrow f_0(i, j) = f(i, j)$  and  $\tilde{f}_\epsilon(i, j) \rightarrow \tilde{f}_0(i, j) = \tilde{f}(i, j)$ , for all pairs  $(i, j) \in T^2$ . From this we obtain the following easy

**Lemma 13.** *Assume that the theoretical probabilities  $P(IJ, KL)$  are known for each four taxon subset  $\{I, J, K, L\} \subset T$  and are used as input for Algorithm 10. Then there exists a constant  $c > 0$  such that Algorithm 10 computes the same results with  $f_\epsilon$  in place of  $f$ , for each  $0 \leq \epsilon < c$ . In other words, if  $S_\epsilon$  is the species tree inferred by Algorithm 10 with  $f_\epsilon$  in place of  $f$ , then the limit  $\lim_{\epsilon \rightarrow 0} S_\epsilon$  exists and is equal to  $S = S_0$ .*

*Proof.* This follows from the fact that  $S$  is a binary tree, that there are only finitely many values of  $f$  and that  $f_\epsilon(i, j)$  depends continuously on  $\epsilon$ .  $\square$

This suggests that for practical applications it will be reasonable to fix  $\epsilon > 0$  ‘small enough’, and run Algorithm 10 with the perturbed empirical depth function  $\tilde{f}_\epsilon$  in place of  $f$  in order to avoid ‘infinite depths’ as in the



discussion at the beginning of this subsection. Hence we obtain the following modification of Algorithm 10.

**Algorithm 14** (Quartet Neighbor Joining, Quartet-NJ). *Input:* A finite set of taxa  $T = \{t_1, \dots, t_n\}$ , a ‘small’ constant  $\epsilon > 0$ , and for each four taxon subset  $\{a, b, c, d\} \subset T$  the relative quartet gene tree frequencies  $r(AB, CD)$ ,  $r(AC, BD)$ ,  $r(AD, BC)$ .

*Output:* An unrooted tree  $\tilde{S}_\epsilon$  on  $T$  which is our estimate for the topology of the species tree  $S$  on  $T$ .

**Algorithm:** Run Algorithm 10 with the depth function  $f(i, j)$  substituted by the empirical depth  $\tilde{f}_\epsilon(i, j)$ . The result of this calculation is  $\tilde{S}_\epsilon$ .  $\square$

Indeed, if we apply this modified algorithm to the problematic situation at the beginning of this subsection, we will obtain (for any choice of  $\epsilon$ ) a fully resolved correct estimate for the species tree  $S$ . Of course, this algorithm has the same complexity as Algorithm 10.

We want to claim that, for sufficiently small values of  $\epsilon$ , quartet-NJ produces a asymptotically correct estimate of the true species tree on  $T$ . To this end, we first prove

**Lemma 15.** *There exists a constant  $c > 0$  such that for all  $0 < \epsilon < c$  the trees  $\tilde{S}_\epsilon$  are equal. In other words, the limit  $\lim_{\epsilon \rightarrow 0} \tilde{S}_\epsilon$  exists.*

*Proof.* Since there are only finitely many cherries to pick, we may restrict our considerations to the picking of the first cherry. We have to distinguish two cases: First assume that there are taxa  $i, j$  such that  $\tilde{f}(i, j) = \infty$ , i.e. the set  $Inf = \{(i, j) | \tilde{f}(i, j) = \infty\}$  is nonempty. If  $\epsilon$  is small enough, then the maximum of the values of  $\tilde{f}_\epsilon$  is attained at one of the pairs in  $Inf$ . At which of those pairs the maximum is attained then only depends on the number of summands in  $\tilde{f}_\epsilon(i, j)$  which approach infinity as  $\epsilon$  tends to zero. This number is clearly independent of  $\epsilon$ , whence the set of potential cherries to pick is independent of  $\epsilon$ .

In the second case, there are no elements in the set  $Inf$ . This means that each of the perturbed empirical depths  $\tilde{f}_\epsilon(i, j)$  approaches the *finite* value  $\tilde{f}(i, j)$  as  $\epsilon$  tends to zero. This again means that, for  $\epsilon$  small enough, the pair which maximizes  $\tilde{f}_\epsilon$  does not depend on  $\epsilon$ .  $\square$

Combining this with Lemma 13 we obtain the desired result.

**Theorem 16** (Statistical consistency of quartet-NJ for small  $\epsilon$ ). *The limit  $\lim_{\epsilon \rightarrow 0}(\tilde{S}_\epsilon)$  exists and is a statistically consistent estimate for the true species tree  $S = S_0$ . In particular, quartet-NJ is statistically consistent.*

*Proof.* The existence of the limit in the theorem is established by Lemma 15. It remains to prove that it is a statistically consistent estimate for  $S$ . Recall that the definition of the function  $\tilde{f}_\epsilon(i, j)$  depends on the relative frequencies  $r(IJ, KL)$ . If we consider the probabilities  $P(IJ, KL)$  known and fixed and moreover fix the pair of taxa  $(i, j)$ , then we may consider the function

$$F(i, j) = \tilde{f}_\epsilon(i, j) - f_\epsilon(i, j) : \mathbb{R}_{>0}^{\binom{n}{4}} \times \mathbb{R} \rightarrow \mathbb{R}, \quad (4.4)$$

depending on the ‘variables’  $r(IJ, KL)$  and  $\epsilon$ . This is a continuous function which vanishes on the line  $\{(P(IJ, KL))_{(IJ, KL)}\} \times \mathbb{R} \subset \mathbb{R}^{\binom{n}{4}} \times \mathbb{R}$ . Thus for every  $\delta$  there exists an open ball centered at  $((P(IJ, KL))_{(IJ, KL)}, 0)$  which is mapped to  $(-\delta, \delta) \subset \mathbb{R}$  by  $F(i, j)$ . In particular, there exists a positive constant  $c > 0$  such that for every  $(i, j)$  we have  $|\tilde{f}_\epsilon(i, j) - f_\epsilon(i, j)| < \delta$  if  $|r(IJ, KL) - P(IJ, KL)| < c$  and  $\epsilon < c$ . If we chose  $\delta$  small enough, we thus may conclude that  $\tilde{S}_\epsilon = S_\epsilon$  for every  $\epsilon < c$  and for all relative frequencies satisfying  $|r(IJ, KL) - P(IJ, KL)| < c$ . Moreover, by Lemma 13 we may assume, by reducing  $c$  if necessary, that  $S_\epsilon = S$  for every  $\epsilon < c$ .

Taking this together we obtain that, provided  $|r(IJ, KL) - P(IJ, KL)| < c$  for every quartet  $(ij, kl)$ ,  $\lim_{\epsilon \rightarrow 0} \tilde{S}_\epsilon = S$ . Since  $r(IJ, KL)$  is a statistically consistent estimate for  $P(IJ, KL)$ , this proves that  $\lim_{\epsilon \rightarrow 0} \tilde{S}_\epsilon$  is a statistically consistent estimate for  $S$ .  $\square$

The question of how to find an appropriate  $\epsilon > 0$  to run the neighbor joining algorithm will of course depend on the special problem instance and is left open here.

### 4.3 Reduction of computational complexity

Here we briefly mention a complexity reduction for the quartet neighbor joining algorithm. Since Algorithm 10 and 14 are completely equivalent in this respect, we formulate this reduction only with the simpler notation of Algorithm 10.

We consider step 2 in Algorithm 10. In our present formulation, this step calculates the depth  $f(i, j)$  by adding up  $O(|T|^2)$ -many quartet weights

for each pair of taxa  $i, j$ . However, most of these quartet weights will not have changed during the last iteration of the algorithm.

Assume that in the previous iteration the cherry  $(i_0, j_0)$  was identified and joined by a new vertex  $x$ . Then for any pair of taxa  $i, j$  which are still present in  $T$  we may calculate the new depth  $f_{new}(i, j)$  from the old one by the formula

$$f_{new}(i, j) = f(i, j) - \sum_{k \in T} (w(ij, i_0k) + w(ij, j_0k)) + \sum_{k \in T} w_{new}(ij, kx).$$

This reduces the complexity of step 2 from  $O(|T|^4)$  to  $O(|T|^3)$ , and hence the overall complexity of Algorithms 10 and 14 are reduced by this modification to  $O(|T|^4)$ .

## 5 Application and simulations

### 5.1 Application to a prokaryote data set

In order to test quartet neighbor joining on real data, Algorithm 14 was applied to a set of protein sequences from nine prokaryotes, among them the two archaea *Archaeoglobus fulgidus* (AF) and *Methanococcus jannaschii* (MJ), as well as the seven bacteria *Aquifex aeolicus* (AQ), *Borrelia burgdorferi* (BB), *Bacillus subtilis* (BS), *Escherichia coli* (EC), *Haemophilus influenzae* (HI), *Mycoplasma genitalium* (MG), and *Synechocystis sp.* (SS).

The choice of organisms follows Teichmann and Mitchison [17], while the multiple sequence alignments were taken out of the dataset used by Cicarelli et al. in [2]. For each of the 28 protein families in the list

1. Ribosomal proteins, small subunits: S2, S3, S5, S7, S8, S9, S11, S12, S13, S15, S17, L1, L3, L5, L6, L11, L13, L14, L15, L16, L22;
2. tRNA-synthetase: Leucyl-, Phenylalanyl-, Seryl-, Valyl;
3. Other: GTPase, DNA-directed RNA polymerase alpha subunit, Pre-protein translocase subunit SecY,

a distance matrix was calculated using BELVU [15] with ‘scoredist’-distance correction (Sonnhammer and Hollich [15]). For each distance matrix, a set of quartet trees was inferred in the classical way by finding, for

each four taxon set  $\{i, j, k, l\}$ , the unrooted quartet  $(IJ, KL)$  which maximizes the value

$$d(I, K) + d(I, L) + d(J, K) + d(J, L) - 2d(I, J) - 2d(K, L),$$

where  $d(-, -)$  denotes the respective entry in the distance matrix.

Algorithm 14 was then run with the parameter  $\epsilon = 10^{-6}, \epsilon = 10^{-9}$ , and  $\epsilon = 10^{-12}$  on the quartet distribution obtained from analyzing all the 28 protein families above, and in a second try on the quartet distributions obtained only by the ribosomal proteins. The resulting tree topologies are depicted in Figures 4 and 5, respectively (The root of these trees is of course not predicted by Algorithm 14. Rather it was placed a posteriori on the branch which separates archaea from bacteria on the unrooted output of the algorithm.) The different choices of  $\epsilon$  did not affect the result in these calculations.

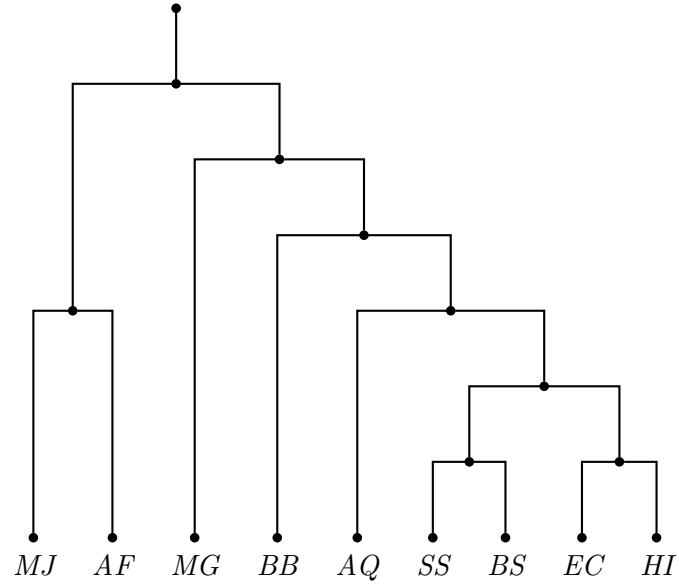


Figure 4: The neighbor joining species tree topology for the nine prokaryotes, using multiple alignments for all 28 protein families.

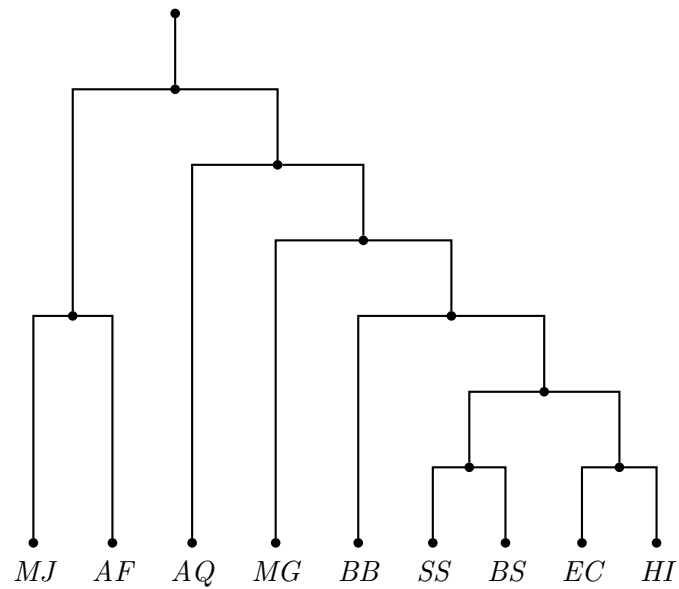


Figure 5: The neighbor joining species tree topology for the nine prokaryotes, using only the multiple alignments of the ribosomal proteins above.

## 5.2 Simulations with Mesquite

As a first test of performance of Algorithm 14 two series of simulations were performed as follows. For a certain choice of a species tree  $S$  on a set of taxa  $T$ , a coalescent process was repeatedly simulated using the coalescence package of Mesquite [10], [11]. Each simulation yielded a set of gene trees on  $T$ , and from these the frequency of each unrooted quartet gene tree with leaves in  $T$  was determined. These frequencies were then submitted to Algorithm 14 and the resulting (unrooted) tree was compared with the (unrooted version of the) species tree  $S$ . We report here the proportion of correct inferences of the unrooted species tree in the different situations.

In the first series of simulations the underlying species tree was the 5-taxon caterpillar tree  $(a, (b, (c, (d, e))))$ , with all internal branch lengths set equal (of course, the length of the pending edges does not have an impact, as we consider only one lineage per taxon). Algorithm 14 was run 1000 times using 5, 10, 20 and 50 sampled gene trees per trial, 500 times using 100 gene trees, 250 times using 200 gene trees, and 100 times using 500 simulated gene trees per trial. The proportion of trials which yielded the correct unrooted species tree topology is reported in Table 1. Note that simulations under the 5-taxon caterpillar tree are also performed by Liu and Yu [7] in order to assess the performance of their ‘neighbor joining algorithm for species trees’. For our choices of branch lengths and sample sizes, the performance of Algorithm 14 seems to be roughly equal to the performance of ‘neighbor joining for species trees’ (Liu and Yu [7], Figure 2).

In a second series of simulations, the underlying species tree was the tree inferred by Algorithm 14 for the nine prokaryotes in Section 5.1, see Figure 4. Again, for different internal branch lengths and different numbers of gene trees per trial, 1000 trials (in the case of 5, 10, 20 and 50 gene trees per trial), and 500 resp. 250 resp. 100 trials (in the case of 100 resp. 200 resp. 500 gene trees per trial) were run, and the proportions of correctly inferred unrooted species tree topologies are reported in Table 2.

Two comments are in order: (1) In fact, for each choice of the parameter  $x$  (and fixed number of gene trees per trial) two simulations were performed. For the first, the length of the branch leading to the cherry formed by MJ and AF was set to  $4x$ , while in a second simulation this branch length was

set to  $2x$ . The differences in the proportions of successful trials are small in most cases (between one and two percent), and, as expected, in most cases the proportion of successful trials was bigger in the first situation.

(2) The number of gene trees (or rather, the number of unrooted quartet gene trees for each 4-taxon subset) used for the reconstruction of the prokaryote species tree in Section 5.1 were 21 and 28, respectively. From Table 2 we see that such a species tree  $S$  is likely to be inferred correctly by Algorithm 14 if its internal branch lengths are around 0.5 (with a probability of more than 90 percent). For branch lengths around 0.2, however, the probability for a correct inference of  $S$  decreases to about 40 to 50 percent. (However, there might still be certain clades on  $S$  which can be detected with high accuracy also for smaller branch lengths.) Clearly, in these considerations we ignore effects such as horizontal gene transfer, for whose existence there is evidence in the case of the nine prokaryotes considered in Section 5.1 for some non-ribosomal proteins (see Teichmann and Mitchison [17]).

$x = \text{internal}$ branch length	5	10	20	50	100	200	500
0.01	0.107	0.100	0.093	0.114	0.124	0.20	0.21
0.05	0.162	0.210	0.230	0.330	0.430	0.67	0.85
0.1	0.285	0.334	0.440	0.662	0.815	0.93	1.00
0.2	0.428	0.580	0.751	0.943	0.990	1.00	1.00
0.5	0.755	0.890	0.980	1.000	1.000		
1.0	0.954	0.992	1.000	1.000			

Table 1: Simulation results for the 5-taxon caterpillar tree ( $a, (b, (c, (d, e)))$ ) with all internal branch lengths set to  $x$  (in coalescent units). Table entries are proportions of trials which yielded the correct unrooted species tree topology. Columns are labelled by the number of simulated gene trees used in each trial.

$x = \text{int.}$ branch l.	5	10	20	25	50	100	200	500
a) 0.1	0.005	0.027	0.068	0.102	0.230	0.320	0.33	0.37
b) 0.1	0.007	0.018	0.064	0.087	0.218	0.282	0.31	0.27
a) 0.2	0.046	0.181	0.406	0.498	0.750	0.876	0.97	1.00
b) 0.2	0.034	0.157	0.385	0.475	0.735	0.902	0.96	1.00
a) 0.5	0.325	0.626	0.899	0.966	0.999	1.000		
b) 0.5	0.303	0.607	0.911	0.957	1.000	1.000		
a) 1.0	0.703	0.873	0.966	0.980	0.999	1.000		
b) 1.0	0.708	0.868	0.969	0.983	0.997	1.000		

Table 2: Simulation results for the prokaryote species tree shown in Figure 4. All internal branch lengths were set to  $x$  (in coalescent units), except for the edge leading from the root to (MJ,AF): the length of this edge was set to a)  $4x$  and to b)  $2x$ , respectively. Columns are labelled by the number of simulated gene trees used in each trial.



## References

- [1] Elizabeth S. Allman, James H. Degnan, and John A. Rhodes. *Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent*. J. Math. Biol. (2011), in press:DOI:10.1007/s00285-010-0355-7.
- [2] Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. *Toward automatic reconstruction of a highly resolved tree of life*. Science (2006), 311(5765):1283-7.
- [3] J. H. Degnan and L. A. Salter. *Gene tree distributions under the coalescent process*. Evolution (2005), 59:24–37.
- [4] J. H. Degnan and N. A. Rosenberg. *Gene tree discordance, phylogenetic inference and the multispecies coalescent*. Trends Ecol. Evol. (2009), 24:332–340.
- [5] L. Salter Kubatko and J. H. Degnan, *Inconsistency of Phylogenetic Estimates from Concatenated Data under Coalescence*. Syst. Biol. (2007) 56(1): 17–24 doi:10.1080/10635150601146041
- [6] L. Liu, L. Yu, and S. V. Edwards. *A maximum pseudo-likelihood approach for estimating species trees under the coalescent model*. BMC Evol. Biol. (2010), 10:302.
- [7] Liang Liu and Lili Yu. *Estimating species trees from unrooted gene trees*. Syst. Biol. (2011) doi: 10.1093/sysbio/syr027.
- [8] W. Maddison. *Gene trees in species trees*. Syst. Biol. (1997), 46(3):523–536.
- [9] W. Maddison and L. Knowles. *Inferring phylogeny despite incomplete lineage sorting*. Syst. Biol. (2006), 55(1):21–30.
- [10] Maddison, W. P. and D.R. Maddison. *Mesquite: a modular system for evolutionary analysis. Version 2.74* (2010) <http://mesquiteproject.org>
- [11] Maddison, W. P. *Coalescence package for Mesquite. Version 2.74*. (2010) <http://mesquiteproject.org>.

- [12] M. Nei. *Molecular Evolutionary Genetics*. Columbia University Press (1987), NY.
- [13] Nei M. *The neighbor-joining method: a new method for reconstructing phylogenetic trees*. Mol. Biol. Evol. (1987) 4(4):406–425.
- [14] L. Pachter and B. Sturmfels. *Algebraic Statistics for Computational Biology*. Cambridge University Press (2005).
- [15] Erik LL Sonnhammer, Volker Hollich, *Scoredist: A simple and robust protein sequence distance estimator*. BMC Bioinformatics (2005), 6:108.
- [16] K. Strimmer and A. von Haeseler. *Quartet Puzzling: A Quartet Maximum-Likelihood Method for Reconstructing Tree Topologies*. Mol. Biol. Evol. (1996), 13(7): 964.
- [17] Sarah A. Teichmann and Graeme Mitchison. *Is There a Phylogenetic Signal in Prokaryote Proteins?* J. Mol. Evol. (1999), 49:98–107.
- [18] C. Than and L. Nakhleh, *Species tree inference by minimizing deep coalescences*. PLoS Computational Biology (2009), 5(9): e1000501.
- [19] C. Than and N. Rosenberg. *Consistency properties of species tree inference by minimizing deep coalescences*. Journal of Computational Biology (2011), 18(1):1–15.
- [20] Y. Yu, T. Warnow, and L. Nakhleh, *Algorithms for MDC-based multi-locus phylogeny inference*. Proceedings of the 15th Annual International Conference on Research in Computational Molecular Biology (RECOMB), LNBI 6577, 531-545, 2011.